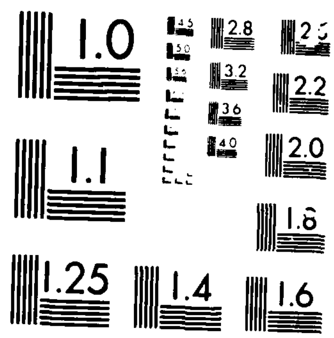
AD-A169 145   STATISTICAL MODELS AND METHODS FOR CLUSTER ANALYSIS AND IMAGE SEGMENTATIO..(U) ILLINOIS UNIV AT CHICAGO CIRCLE DEPT OF INFORMATION AND DECIS..  S L SCLOVE 15 MAR 86   1/1

UNCLASSIFIED   UIC/DQM/ARO-85-2 ARO-19085.12-MA   F/G 17/8   NL


1.0
2.8
3.2
3.6
4.0
2.2
2.0
1.1
1.8
1.25
1.4
1.6

ARO 19085.12-MA

AD-A169 145

FINAL REPORT:

ARMY RESEARCH OFFICE CONTRACT DAAG29-82-K-0155

# STATISTICAL MODELS AND METHODS FOR CLUSTER ANALYSIS AND IMAGE SEGMENTATION

Stanley L. Sclove

TECHNICAL REPORT NO. UIC/DQM/ARO 85-2
March 15, 1986

PREPARED FOR THE
ARMY RESEARCH OFFICE
UNDER
CONTRACT DAAG29-82-K-0155

Statistical Models and Methods for
Cluster Analysis and Image Segmentation

Principal Investigator: Stanley L. Sclove



INFORMATION & DECISION SCIENCES DEPARTMENT
COLLEGE OF BUSINESS ADMINISTRATION
UNIVERSITY OF ILLINOIS AT CHICAGO
BOX 4348, CHICAGO, IL 60680

3/30/86

# STATISTICAL MODELS AND METHODS FOR CLUSTER ANALYSIS AND IMAGE SEGMENTATION

Stanley L. Sclove

Department of Information & Decision Sciences
University of Illinois at Chicago

## CONTENTS



A-1

Final Report:
ARO Contract DAAG29-82-K-0155

# Statistical Models and Methods for CLUSTER ANALYSIS AND IMAGE SEGMENTATION

by

Stanley L. Sclove
Department of Information and Decision Sciences
College of Business Administration
University of Illinois at Chicago

## ABSTRACT

Clustering of individuals, segmentation of time series and segmentation of numerical images can all be considered as labeling problems, for each can be described in terms of pairs $(x_t, g_t)$, $t = 1,2,\ldots,n$, where $x_t$ is the observation at instance t and $g_t$ is the unobservable "label" of instance t. The labels are to be estimated, along with any unspecified distributional parameters. In cluster analysis the values of t are the individuals (cases) observed and the x's are independent. In time series the values of t are time instants and there is temporal correlation. In numerical image segmentation the values of t denote picture elements (pixels) and spatial correlation between neighboring pixels can be utilized. The idea in segmentation is that signals and time series often are not homogeneous but rather are generated by mechanisms or processes with various phases. Similarly, images are not homogeneous but contain various objects. "Segmentation" is a process of attempting to recover automatically the phases or objects. The present report summarizes the work done on these problems under ARO Contract DAAG29-82-K-0155.

Key words and phrases:
statistical pattern recognition; classification;
temporal correlation, spatial correlation;
optimization by relaxation method.

## 1. Introduction

The research reported here relates to cluster analysis and to statistical processing of time series and digitized images. This report is a summary of work performed under ARO Contract DAAG29-82-K-0155 (6/15/82 - 6/15/85): Statistical Models and Methods for Cluster Analysis and Image Segmentation. The type of datasets to which the techniques developed are applicable include: signals such as radar and sonar; economic and bio-medical time series; time series arising from quality assurance acceptance sampling by attributes or variables; and digital images which can result from various sources, including bio-medical imagery, infrared imagery obtained by smart munitions, and multispectral data obtained by satellite. The problems addressed are those of clustering, and segmentation of time series and images.

The work involves the further development of algorithms for clustering large, multidimensional datasets and for segmentation of time series and digital images. The algorithms are based on maximum likelihood estimation in distribution-mixture models. In the context of these mixture models clustering is construed as estimation of unobserved labels. An observation's label, were it observable, would tell from which mixture component the observation arose. Image segmentation is also considered as a labeling problem. Throughout the work there is an attempt to apply model-selection criteria to the decision as to an appropriate number of clusters or classes of segment.

Software development is an important aspect of such a project. The algorithms developed are programmed in FORTRAN.

Some of the ideas developed in the project have already been published; see Sclove (1983a,b,c; 1984a) and Bozdogan and Sclove (1984).

The organization of the present paper is as follows: Section 2 concerns cluster analysis; in this section there is some general discussion of model-selection criteria and a digression to mention some ideas concerning clustering of variables. Section 3 summarizes some of the results on time-series segmentation, and results on image segmentation are discussed in Section 4.

## 2. Cluster analysis

Background. The mixture model for the clustering problem treats the sample as having arisen from a mixture of several (k) distributions. This is the approach put forth in (Sclove 1977). The research problem set there was, at least in part, to see whether the ISODATA (Ball and Hall, 1967) and K-MEANS (MacQueen, 1967) algorithms could be interpreted as mathematical-statistical estimation schemes in some model for the clustering problem. That is, did there exist a model for the clustering problem, and an estimation method in that model, such that ISODATA and K-MEANS corresponded to that method applied to that model? The answer, provided in (Sclove 1977), was affirmative; this will be explained below, but first let us briefly define ISODATA and K-MEANS.

The "isodata" scheme proceeds as follows. One starts with tentative estimates of cluster means as seed points for the clusters and assigns each observation to the mean to which it is closest. The cluster means are then re-estimated, and one loops through the data

again, reassigning the observations. Etc. In the K-MEANS algorithm, the seed points are updated immediately after each observation is tentatively classified. In (Sclove 1977) it was shown that these algorithms correspond to iterative maximum likelihood estimation in a type of mixture model for the clustering problem, where the component distributions are multivariate normal.

This clustering can be done for various values of k, the number of clusters. Figures of merit can be used to choose the best k. Model-selection criteria can be used as figures of merit.

### 2.1. Model-selection criteria

In the context of a mixture model, choice of the number of clusters $k$ can be viewed as a model-selection problem. However, at least in the case of clustering individuals, existing model-selection criteria have to be modified, as they depend upon (regularity) assumptions that are not always met in mixture models for clustering individuals.

In any case, let us review some of the existing model-selection criteria. Consider, then, a problem of choosing from among several models, indexed by $k$ $(k = 1,2,\ldots,K)$. Let $L(k)$ be the likelihood, given the k-th model. Various model-selection criteria taking the form

$$-2 \log(\max L(k)) + a(n)m(k) + b(k), \qquad (1)$$

have been developed in relatively recent years. Here $n$ is the sample size, log denotes the natural logarithm, $\max L(k)$ denotes the maximum of the likelihood over the parameters, and $m(k)$ is the number of independent parameters in the k-th model. For a given criterion, $a(n)$ is the cost of fitting an additional parameter and $b(k)$ is an

additional term depending upon the criterion and the model k. One chooses the model k for which the value of the criterion being used is smallest.

Akaike (see, e.g., Akaike 1973, 1974, 1981) developed such a criterion as an (heuristic) estimate of the expected entropy (Kullback-Leibler information). Akaike's information criterion (AIC) is of the form (1) with

$$a(n) = 2 \text{ for all } n, \quad b(k) = 0 \quad \text{(AIC)}. \tag{2}$$

Schwarz (1978), working from a Bayesian viewpoint, obtained a criterion of the form (1) with

$$a(n) = \log n, \quad b(k) = 0 \quad \text{(Schwarz' criterion)}. \tag{3}$$

Since, for $n$ greater than 8, $\log n$ exceeds 2, it follows that Schwarz' criterion favors models with fewer parameters than does Akaike's.

Noting that AIC has $a(n)$ a constant function of n, namely 2, various researchers, including Kashyap (1982) and Schwarz (1978) have mentioned that AIC is not consistent; $a(n)$ needs to depend upon n.

Kashyap (1982), also working from a Bayesian approach, took the asymptotic expansion of the logarithm of the posterior probabilities a term further than did Schwarz and obtained the criterion of the form (1) given by

$$a(n) = \log n, \quad b(k) = \log(\det B(k)) \quad \text{(Kashyap's criterion)}, \tag{4}$$

where det denotes the determinant and $B(k)$ is the negative of the matrix of second partials of $\log L(k)$, evaluated at the maximum likelihood estimates. In Gaussian linear models this is the covariance matrix of the maximum likelihood estimates of the regression coefficients; in general, the expectation of $B(k)$, evaluated at the

true parameter values, is Fisher's information matrix. Since Kashyap's criterion is based on reasoning similar to Schwarz', but contains an extra term, it may perform better. [Further comments on model-selection criteria are made in Sclove (1983d).]

2.2. Multi-sample clustering

The problem of multi-sample clustering, the grouping of samples, is treated in Bozdogan and Sclove (1984). The situation is the K-sample problem (one-way analysis of variance), with an emphasis on grouping the samples into fewer than K clusters. The use of model-selection criteria in this context can provide an alternative to multiple-comparison procedures. Use of model-selection criteria avoids the difficult choice of levels of significance in such problems. Model-selection criteria can also be used in this context to decide whether or not to assume a common covariance matrix. Kashyap's criterion could be evaluated and used for these problems.

2.3. Clustering of individuals

Schwarz' and Kashyap's criteria could be calculated for the problem of clustering individuals according to Wolfe's (1970) mixture-model clustering approach and incorporated into computer programs for clustering. The values of the criteria can be used heuristically as figures of merit for alternative models, but in order to be rigorously applied the model-selection criteria need to be modified since their derivation involves an assumption of nonsingularity of the information matrix. However, note in this regard a potential advantage of model-selection criteria over a hypothesis-testing approach in this and similar situations. Model-selection criteria require nonsingularity of the information

matrix only for each fixed model k. The testing approach runs into difficulties because of nonsingularity of the matrix at the boundary between the null and alternative hypotheses (i.e., at the boundary between models).

2.4. Clustering of variables

The clustering of variables can also be viewed as a model-selection problem. For example, whether and how to cluster multinormal variables depends upon which covariances may be assumed to be zero; the possible patterns of zeros among the covariances are separate models, a figure of merit for which is provided by a suitable model-selection criterion. This idea is to be further developed.

## 3. Time-series segmentation

As mentioned above, a model for clustering or segmentation is given by assuming that each instance of observation, t, gives rise not only to an observation $x_t$ but also to a label, $g_t$, equal to 1, 2, ..., or k, where k is the number of classes of segment. Model-selection criteria are used to estimate k. In the context of this model, segmentation is merely estimation of the labels. Sclove (1983b,c; 1984a) treats the problem by modeling the label process as a Markov cha'n. An algorithm and computer programs are discussed; numerical examples are given.

The model involves three sets of parameters: the distributional parameters (e.g., means and covariance matrices), the labels, and the transition probabilities between labels.

The algorithm is a relaxation method, similar to the EM algorithm. The estimation step consists of maximum-likelihood estimation of the

distributional parameters, for tentatively fixed values of the labels and transition probabilities. The maximization step consists of maximizing the likelihood over the labels and transition probabilities, for tentatively fixed values of the distributional parameters.

As developed so far, the algorithm is a forward algorithm, classifying $x_2$ after $x_1$, $x_3$ after $x_2$ and $x_1$, etc. It is suitable for sequential operation in real time, but it is non-optimal in other modes of operation. Its performance could possibly be improved by a backcasting technique analogous to that in Box and Jenkins (1976) and by application of the Viterbi algorithm (Forney 1973), which is a recursive optimal solution to the problem of estimating the state sequence of a discrete-time finite state Markov process; it is applicable here because this is what we have at each stage when the distributional parameters and transition probabilities are tentatively fixed and the labels are to be estimated.

Further, the parameter-estimation step of the algorithm can be improved. The estimation implemented in the existing algorithm leads to estimates that are biased (even asymptotically). (See, e.g., Bryant and Williamson 1978.) This bias may be viewed as due to the truncation resulting from the algorithm. The estimation could be modified by doing it in a Bayesian manner, e.g., estimate the mean of Class A as

$$\sum_{t=1}^{n} x_t \Pr(a|x_{t\delta}) / \sum_{t=1}^{n} \Pr(a|x_t)$$

(In this expression, $\Pr(a|x)$ can be replaced by $\Pr(x|a)$ since $\Pr(a)/f(x)$ will cancel out.) This modification in the parameter-estimation step can be important. For, in this estimate,

all the observations play a role, whether labeled as "Class A" or otherwise, so that at least some of the bias incurred by using only the "a" observations will be removed by allowing all of the observations to enter.

The work done to date is explicit only for the case in which the class-conditional processes consist of independent, identically distributed random variables. The work is to be extended to other, often more realistic cases, such as that of autoregression within segments.

## 4. Image segmentation

Similar ideas are applied to digital images in Sclove (1983a;1984a). Here the label process is modeled as a Markov random field. The same improvements made in the time-series context will be carried over to the two-dimensional, image-processing context. For example, computer experiments (Sclove 1984b) with the existing algorithm have shown it to be successful, even in finding small targets. However, at the same time, these experiments have shown the importance of some such modification as backcasting, as mentioned in connection with time series, to eliminate anomalous border effects.

Extension of the existing work to two-dimensional autoregressions within segments will yield algorithms that may detect textures.

## References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Proceedings of the 2nd International Symposium on Information Theory, 267-281. Akademia Kiado, Budapest.

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control 6, 716-723.

Akaike, H. (1981). Likelihood of a model and information criteria. Journal of Econometrics 16, 3-14.

Ball, G. H., and Hall, D. J. (1967). A clustering technique for summarizing multivariate data. Behavioral Science 12, 153-155.

Box, G.E.P., and Jenkins, G.M. (1976). Time Series Analysis: Forecasting and Control, rev. ed. John Wiley & Sons, New York.

Bozdogan, Hamparsum, and Sclove, Stanley L. (1984). Multi-sample cluster analysis using Akaike's information criterion. Annals of the Institute of Statistical Mathematics 36, 163-180.

Bryant, P., and Williamson, J. A. (1978). Asymptotic behaviour of classification maximum likelihood estimates. Biometrika 65, 273-281.

Forney, G. David, Jr. (1973). The Viterbi algorithm. Proceedings of the IEEE, Vol. 61, 268-278.

Kashyap, R. L. (1982). Optimal choice of AR and MA parts in autoregressive moving average models. IEEE Transactions on Pattern Analysis and Machine Intelligence 4, 99-104.

MacQueen, J. (1966). Some methods for classification and analysis of multivariate observations. Pages 281-297 in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1. University of California Press, Los Angeles and Berkeley.

Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics 6, 461-464.

Sclove, Stanley L. (1977). Population mixture models and clustering algorithms. Communications in Statistics(A) 6, 417-434.

Sclove, Stanley L. (1983a). Time-series segmentation: a model and a method. Information Sciences 29, 7-25.

Sclove, Stanley L. (1983b). Application of the conditional population-mixture model to image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 5, 428-433.

Sclove, Stanley L. (1983c). On segmentation of time series. In Studies in Econometrics, Time Series, and Multivariate Statistics (S. Karlin, T. Amemiya, and L. Goodman, eds.), Academic Press, 1983, 311-330.

Sclove, Stanley L. (1983d). Use of model-selection criteria in clustering and segmentation of time series and digital images. Contributed paper, 44th Session of the International Statistical Institute, Madrid, 9/12-22/83.

*******************************************************************

Sclove, Stanley L. (1984a). On segmentation of time series and images in the signal detection and remote sensing contexts. Pages 421-434 in Statistical Signal Processing (Edw. J. Wegman and James G. Smith, eds.), Marcel Dekker, Inc., New York.

Sclove, Stanley L. (1984b). On segmentation of signals, time series, and Images. Pages 267-289 in Proceedings of the 30th Conference on Design of Experiments in Army Research, Development and Testing, Las Cruces, NM, 10/15-19/84 (ARO Report 85-2).

Sclove, Stanley L. (1986). Statistical models and methods for cluster analysis and segmentation. To appear in Proceedings of the 31st Conference on Design of Experiments in Army Research, Development and Testing, Madison, Wisc., 10/21-25/85.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Research 5, 329-350.

## List of Project Personnel

Principal Investigator:

Stanley L. Sclove

Professor
Department of Information & Decision Sciences
(former name: Department of Quantitative Methods)
College of Business Administration
University of Illinois at Chicago

Associate Investigator:

Hamparsum Bozdogan

Assistant Professor
Department of Mathematics
University of Virginia
(formerly Assistant Professor,
Department of Quantitative Methods,
University of Illinois at Chicago)

***********************************************************************

TECHNICAL REPORTS

ARMY RESEARCH OFFICE CONTRACT DAAG29-82-K-0155

with the University of Illinois at Chicago

Statistical Models and Methods for
Cluster Analysis and Image Segmentation

Principal Investigator: Stanley L. Sclove

No. A82-1. Stanley L. Sclove. "Application of the Conditional Population-Mixture Model to Image Segmentation." 8/15/82

No. A82-2. Hamparsum Bozdogan and Stanley L. Sclove. "Multi-sample Cluster Analysis using Akaike's Information Criterion." 12/20/82

No. A82-3. Stanley L. Sclove. "Time-Series Segmentation: a Model and a Method." 12/22/82

No. A83-1. Hamparsum Bozdogan. "Determining the Number of Component Clusters in the Standard Multivariate Normal Mixture Model using Model-Selection Criteria." 6/16/83

No. A83-2. Stanley L. Sclove. "On Segmentation of Digital Images using Spatial and Contextual Information via a Two-Dimensional Markov Model." Working Paper: 4/11/83; Technical Report: 12/6/83

No. A83-3. Stanley L. Sclove. "Use of Model-Selection Criteria in Clustering and Segmentation of Time Series and Digital Images." 5/5/83

No. A84-1. Stanley L. Sclove. "Pattern Recognition." 2/1/84

No. A84-2. Stanley L. Sclove. "On Segmentation of Signals, Time Series, and Images." 3/1/85. (Presented to the 30th Conference on Design of Experiments in Army Research, Development, and Testing, 10/17-19/84.)

No. A84-3. Hamparsum Bozdogan. "Multi-Sample Cluster Analysis as an Alternative to Multiple Comparison Procedures." 7/20/84

No. A85-1. Stanley L. Sclove. "Statistical Models and Methods for Cluster Analysis and Segmentation." 3/15/86. (Presented to the 31st Conference on Design of Experiments in Army Research, Development, and Testing, 10/21-25/85.)

3/29/86

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

AD-A169145

## REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: ARO 19085.12-MA
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): Final Report, ARO Contract DAAG29-82-K-0155: Statistical Models and Methods for Cluster Analysis and Image Segmentation
5. TYPE OF REPORT & PERIOD COVERED: Final 6/15/82-6/15/85
6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Stanley L. Sclove
8. CONTRACT OR GRANT NUMBER(s): DAAG29-82-K-0155
9. PERFORMING ORGANIZATION NAME AND ADDRESS: University of Illinois at Chicago, Box 4348, Chicago, IL 60680
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office, Post Office Box 12211, Research Triangle Park, NC 27709
12. REPORT DATE: March 15, 1986
13. NUMBER OF PAGES: i + 13
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

statistical pattern recognition, classification; temporal correlation, spatial correlation; optimization by relaxation method

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

(PLEASE REFER TO CONTINUATION SHEET.)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Clustering of individuals, segmentation of time series and segmentation of numerical images can all be considered as labeling problems, for each can be described in terms of pairs $(x_t, g_t)$, $t=1,2,\ldots,n$, where $x_t$ is the observation at instance t and $g_t$ is the unobservable "label" of instance t. The labels are to be estimated, along with any unspecified distributional parameters. In cluster analysis the values of t are the individuals (cases) observed and the x's are independent. In time series the values of t are time instants and there is temporal correlation. In numerical image segmentation the values of t denote pciture elements (pixels) and spatial correlation between neighboring pixels can be utilized. The idea in segmentation is that signals and time series often are not homogeneous but rather are generated by mechanisms or processes with various phases. Similarly, images are not homogeneous but contain various objects. "Segmentation" is a process of attempting to recover automatically the phases or objects. The present report summarizes the work done on these problems under ARO Contract DAAG29-82-K-0155.

END

DTIC

7-86